A compilation of soybean ESTs: generation and analysis

Randy Shoemaker, Paul Keim, Lila Vodkin, Ernest Retzel, Sandra W. Clifton, Robert Waterston, David Smoller, Virginia Coryell, Anupama Khanna, John Erpelding, Xiaowu Gai, Volker Brendel, Christina Raph-Schmidt, E.G. Shoop, C.J. Vielweber, Matt Schmatz, Deana Pape, Yvette Bowers, Brenda Theising, John Martin, Michael Dante, Todd Wylie, and Cheryl Granger

Abstract: Whole-genome sequencing is fundamental to understanding the genetic composition of an organism. Given the size and complexity of the soybean genome, an alternative approach is targeted random-gene sequencing, which provides an immediate and productive method of gene discovery. In this study, more than 120 000 soybean expressed sequence tags (ESTs) generated from more than 50 cDNA libraries were evaluated. These ESTs coalesced into 16 928 contigs and 17 336 singletons. On average, each contig was composed of 6 ESTs and spanned 788 bases. The average sequence length submitted to dbEST was 414 bases. Using only those libraries generating more than 800 ESTs each and only those contigs with 10 or more ESTs each, correlated patterns of gene expression among libraries and genes were discerned. Two-dimensional qualitative representations of contig and library similarities were generated based on expression profiles. Genes with similar expression patterns and, potentially, similar functions were identified. These studies provide a rich source of publicly available gene sequences as well as valuable insight into the structure, function, and evolution of a model crop legume genome.

Key words: *Glycine max*, genome sequencing, functional genomics.

Résumé : Le séquençage de génomes complets est requis pour la compréhension de la composition génétique d'un organisme. Étant donné la taille et la complexité du génome du soja, le séquençage ciblé de gènes choisis au hasard constitue une approche alternative qui procure une méthode rapide et productive en vue d'identifier des gènes. Dans le cadre du présent travail, plus de 120 000 étiquettes de séquences exprimées (EST) du soja, provenant de 50 banques d'ADNc, ont été évaluées. Ces EST formaient 16 928 contigs et 17 336 étiquettes uniques. En moyenne, chaque contig comprenait 6 EST et totalisait 788 nucléotides. La taille moyenne des séquences soumises à dbEST était de 414 pb. En se limitant aux seules banques qui avaient contribué au moins 800 EST et aux seuls contigs comptant au moins 10 EST, des corrélations de l'expression génique ont pu être observées entre les banques et parmi des gènes. Des représentations qualitatives bidimensionnelles de la similarité des banques et des contigs ont été générées à partir des profils d'expression. Des gènes montrant des profils d'expression semblables, et potentiellement des fonctions similaires, ont été identifiés. Ces études contribuent un vaste éventail de séquences de gènes, dans le domaine publique, et apportent un éclairage nouveau sur la structure, la fonction et l'évolution du génome d'une légumineuse modèle.

Mots clés : *Glycine max*, séquençage génomique, génomique fonctionnelle.

[Traduit par la Rédaction]

Received 20 July 2001. Accepted 16 November 2001. Published on the NRC Research Press Web site at http://genome.nrc.ca on 25 February 2002.

Corresponding Editor: J.P. Gustafson.

R. Shoemaker,[1] J. Erpelding, and C. Granger. USDA–ARS, Corn Insect and Crop Genetics Research Unit, and Department of Agronomy, Iowa State University, Ames, IA 50011, U.S.A.
P. Keim and V. Coryell. Department of Biology, Northern Arizona University, Flagstaff, AZ 86011-5640, U.S.A.
L. Vodkin and A. Khanna. Department of Crop Sciences, University of Illinois, Urbana, IL 61821, U.S.A.
E. Retzel, C. Raph-Schmidt, and E.G. Shoop. Center for Computational Genomics and Bioinformatics, University of Minnesota, Minneapolis, MN 55455-0312, U.S.A.
S.W. Clifton, R. Waterston, D. Pape, Y. Bowers, B. Theising, J. Martin, M. Dante, and T. Wylie. Washington University Genome Sequencing Center, Washington University, St. Louis, MO 63130, U.S.A.
D. Smoller, C.J. Vielweber, and M. Schmatz. IncyteGenomics, St. Louis, MO 63134, U.S.A.
X. Gai and V. Brendel. Department of Zoology and Genetics, Iowa State University, Ames, IA 50011, U.S.A.

[1]Corresponding author (e-mail: rcsshoe@iastate.edu).

 DOI: 10.1139/G01-150

Introduction

Soybean (*Glycine max* L. Merr.) has emerged as a model crop system because of its densely saturated genetic map (Cregan et al. 1999), a well-developed genetic transformation system (Clemente et al. 2000; Xing et al. 2000; Zhang et al. 1999), and the growing number of genetic tools applicable to this biological system (Shoemaker 1999). It is the number one oilseed crop in the world and a multibillion-dollar crop for the United States (Riley 1999; SoyStats 2000). The soybean genome contains approximately 1.12×10^9 base pairs per haploid genome, making it about half the size of the maize (*Zea mays*) genome, but about seven and a half times larger than the *Arabidopsis* genome (Arumuganathan and Earle 1991). About 40–60% of the soybean genome consists of repetitive sequences (Goldberg 1978; Gurley et al. 1979).

Genome sequencing is the cornerstone of functional analysis and is fundamental to understanding the genetic composition of an organism. Whole-genome sequencing is currently underway for rice (*Oryza sativa*) and is essentially complete for the five *Arabidopsis* chromosomes with gap filling and annotation now in progress (Theologis et al. 2000; Salanoubat et al. 2000; Tabata et al. 2000; Mayer et al. 1999; Lin et al. 1999). Because of the size and complexity of the soybean genome, it is unlikely that the entire genome will be sequenced in the near future. However, the composition and general structure of the soybean genome has been estimated by the genomic sampling of nearly 2700 genomic sequences from more than 600 mapped loci (Marek et al. 2001). Using other sampling approaches, genes with differential expression in various organs and tissues, as well as in different temporal, environmental, and biological milieux, may be identified. This programmed clustering of the data provides essential information about the regulation of genes and the metabolic regulation of the organism.

For nearly two decades, random sequencing of gene transcripts has been recognized as a simple and efficient method of identification of many of the expressed genes in an organism (Putney et al. 1983; Adams et al. 1991). These expressed sequence tags (ESTs) have become a valuable and efficient method for gene discovery (Sterky et al. 1998; Hillier et al. 1996; Marra et al. 1999*a*). When sampling is truly random, the frequency of any given EST is related to differentials in gene expression (Adams et al. 1995; Manger et al. 1998; Tanabe et al. 1999; Ewing et al. 1999). ESTs also provide an opportunity to study gene evolution, to make comparative analyses between genera, and when coupled with genetic mapping, to identify candidate genes for important biological processes and phenotypes (Hatey et al. 1998).

Global, multi-tissue EST projects have been reported for *Arabidopsis* (Delseny et al. 1997), rice (Ewing et al. 1999), and maize (Fernandes et al. 2002). More specialized, tissue-specific EST projects have been reported for root hair enriched *Medicago truncatula* tissue (Covitz et al. 1998), flower buds of Chinese Cabbage (*Brassica campestris* subsp. *pekinesis*) (Lim et al. 1996), and wood-forming tissues of poplar (*Populus* spp.) (Sterky et al. 1998). To date, no reports exist of global, multi-tissue EST analyses from a major crop legume.

Here we report on the progress of a public EST project for soybean. At this time, we have analyzed more than 120 000 ESTs generated from more than 50 cDNA libraries representing a wide range of organs, developmental stages, genotypes, and environmental conditions. Through these analyses, we have been able to demonstrate correlated patterns of gene expression across cDNA libraries. As a result, we have been able to develop gene expression profiles across libraries and group cDNA libraries with similar EST composition and genes with similar expression patterns and potentially similar functions. These studies provide a large resource of publicly available genes and gene sequences and provide valuable insight into the structure, function, and evolution of a model crop legume.

Methods and materials

Library construction

cDNA libraries were prepared from tissues representing a wide range of plant developmental stages, organs, genotypes, and biotic and abiotic challenges. Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a *Xho*I restriction site. After *Xho*I digestion, *Eco*RI adapters were ligated to blunt-ended cDNA fragments. cDNAs were directionally cloned into the *Eco*RI–*Xho*I restriction site of pBlueScript II SK+ vector (Stratagene, La Jolla, Calif.). Some libraries were prepared from DNA synthesized from mRNA using a poly(dT) primer with a *Not*I restriction site. *Sal*I linker adapters were ligated to the blunt-ended cDNA fragments followed by *Not*I digestion. The cDNA fragments were then directionally cloned into the *Not*I-*Sal*I site of the pSPORT 1 vector (Invitrogen Life Technologies, Carlsbad, Calif.). Other libraries were directionally cloned into the *Eco*RI–*Not*I site of pT7T3-Pac (Pharmacia, Peapack, N.J.), and in the *Eco*RI–*Xho*I site of pBlueScript II SK+ (Stratagene). The ligated cDNA fragments were then transformed into DH10B (Gibco BRL, La Jolla, Calif.) or XL 10-Gold (Stratagene, La Jolla, Calif.) host cells.

Clone preparation and sequencing

After plating the bacteria containing the cDNA libraries, colonies were picked robotically and assigned a unique identifier. Glycerol stocks were prepared in 384-well format. Stocks of the cDNA clones to be end-sequenced were sent to the Washington University Genome Sequencing Center (Washington University, St. Louis, Mo.) in 384-well format. Clones were transferred into 96-well blocks and incubated at 37°C for 24 h while shaken at 297 rpm in an incubator shaker. Clones were processed according to Marra et al. (1999*b*) using a high-throughput 96-well microwave protocol. Dideoxy terminator sequencing reactions were conducted as described in Hillier et al. (1996).

cDNA size determination

Electrophoresis of restriction enzyme-released cDNA inserts was conducted on 2.0% agarose gels containing 48 μg ethidium bromide/L. Gels were cased into trays containing slots for four combs, each forming 24 sample wells with two flanking wells for size standard markers. Thus, each gel accommodated 96 samples. Electrophoresis was conducted in 1× Tris-acetate–EDTA (TAE; 40mM Tris-acetate – 1 mM EDTA) plus 20 μL of ethidium bromide (10 mg/mL) at 2.1

V/cm for approximately three hours. Gels were visualized using a FluorImager SI (Molecular Dynamics, Sunnyvale, Calif.). Sizes of restriction fragments were obtained interactively using FragmeNT Analysis version 1.1 (Molecular Dynamics) and comparison with the DNA size markers (Marker IV, Boehringer-Mannheim).

Sequence analysis and annotation

Bases were initially called from the gel trace files using the PHRED basecaller (Ewing et al. 1998) and subsequently trimmed of vector sequence and evaluated for quality. A dbEST format file was automatically generated for each trace. Basic information such as library name, primer used, directionality of clone, etc. was included in this file. BLASTN (Altschul et al. 1997) was run on trimmed sequences against a sequencing vector database to verify that the expected sequencing vector was removed. Repeat sequences were masked and potential contaminants removed by screening the sequences against databases for structural RNA sequence, bacterial sequence, mouse and human mitochondrial sequence, plant mitochondrial sequence, plant chloroplast DNA sequence, and plant rRNA sequence. BLASTX (Altschul et al. 1997) was run on trimmed sequences against GenBank's non-redundant (nr) protein database to detect similarities to known genes. Only the single-best match was reported. After completing these analyses, the dbEST-format submission file was submitted to dbEST.

Contig assembly

ESTs from individual *Glycine* spp. libraries were assembled into contigs using the default parameters of CAP3 (Huang and Madan 1999). Incorporation of ESTs into a contig required at least 95% sequence identity and a minimum of a 40-bp overlap. The CAP3 assembly program was also used to assemble the entire *Glycine max* EST database. However, because of the large size of the database (>121 000 ESTs) a stepwise protocol was implemented by which the most abundant ESTs were assembled into contigs and the remaining ESTs were then added to this assembly using the ZmDBAssembler (http://www.zmdb.iastate.edu/zmdb/EST/assembly.html; this script determines preliminary clustering on the basis of sequence similarity reported by BLASTN (Altschul et al. 1997) and derives final contigs by application of CAP3 on the preliminary clusters). Consensus sequences of the resulting contigs and singletons were used to obtain nearest matches in SWISSPROT using BLASTX (Altschul et al. 1997).

Cluster analysis

Both contigs and libraries were clustered with respect to expression profiles as described in Ewing et al. (1999). To diminish a statistical bias, libraries containing less than 800 ESTs and contigs containing less than 10 ESTs were removed from the cluster analysis. In addition, nonrandomly sampled or reracked libraries (Gm-r1021, -r1030, and -r1070), as well as libraries classified as "other", were excluded from the clustering analysis. Pearson's correlation coefficients (r) for both interlibrary and intercontig profile comparisons were determined for the remaining 1763 contigs and 42 libraries as described in Ewing et al. (1999); however, in our analysis, percentage data (counts/library size) were used instead of raw EST counts. Euclidean distances (d) between vectors of correlation coefficients were determined as described in Ewing et al. (1999) and dendrograms were constructed by the UPGMA method (Phylip 3.5c; Felsenstein 1989). These dendrograms were used to reorder the original data such that libraries and contigs with similar expression profiles grouped together. To facilitate qualitative visualization, contig expression profiles were normalized with respect to contig size. Normalized contig profiles that had the lowest ($d < 18$) and highest ($d > 85$) distances compared with an evenly distributed profile (i.e., approximately 2.4% expression in each of the 42 libraries) were described as the most uniformly and non-uniformly expressed contigs, respectively. Identification of the most highly expressed contigs was based on average percent expression over all libraries.

Results and discussion

Forty-nine directionally cloned cDNA libraries were constructed representing a wide range of soybean genotypes, organs, and developmental stages (Table 1). These libraries were randomly sampled and sequenced at the 5′ end and a total of 104 985 EST sequences were submitted to dbEST. All of our analyses involving expression profiles and clustering were based on this random sample. Another library (Gm-r1030; 3073 ESTs) represented a reselection of clones after filter normalization to identify low-copy number cDNAs. Two additional libraries (Gm-r1021 and Gm-r1070; 3274 and 7360 ESTs, respectively) consisting of 3′ end sequences were generated for the purposes of developing a unigene set of clones and will be detailed in another report. All other soybean ESTs from a number of independent dbEST submissions were included in a general category of "other", but were not considered in further clustering analyses (Table 1). Thus, more than 120 000 soybean ESTs are characterized in Table 1, and approximately 105 000 met our criteria for consideration for clustering analysis.

After editing, the average length of sequences submitted to dbEST was 414 bases. Based on stringent sequence similarity, the set of ESTs was assembled into 16 928 contigs and 17 336 singletons, together comprising a set of 34 264 unique gene fragments ("unigenes"). The depth of each library is estimated by its sequence redundancy (number of contigs plus singletons divided by the number of ESTs in this library). This value is influenced by the size of the library (Fig. 1A). As more cDNAs are randomly selected from a library, more redundancy is observed. Among the 49 libraries, the average unigene composition was 75% (Table 1), but ranged from 44 (cotyledons) to 93% (apical shoots). This range is not unexpected given the high redundancy of gene messages in developing cotyledons and the complex gene expression expected from meristematic tissue. Neither the average contig size (number of ESTs in a contig) nor the average contig length are appreciably affected by the number of ESTs sampled from each library (Figs. 1B and 1C). These values suggest that EST sampling for each library was mostly far from being exhaustive, yielding a high gene-discovery rate.

Including the 3′ sequences and considering all libraries in the assembly, the average size and length of each contig was six ESTs and 788 bases, respectively (Table 1). Many

Table 1. Contig assembly information for individual libraries as well as for the entire EST set.

Library	Genotype	Tissue	No. of ESTs	No. of contigs	No. of singlets	% unigenes	Size	Length
Gm-c1003	'Williams'	Cotyledon, mid-maturation	97	17	60	79	2.2	562
Gm-c1004	'Williams'	Root, 8 5day-old	7 674	1 088	3 623	61	3.7	568
Gm-c1005	'Williams 82'	Apical shoot, 9–10 days old	193	12	167	93	2.2	601
Gm-c1006	'Williams 82'	Stem, 2–3 weeks old	142	35	56	64	2.5	455
Gm-c1007	'Williams'	Cotyledon, mature	2 366	138	908	44	10.6	692
Gm-c1008	'Williams'	Pod, 2 cm	1 950	181	1 271	74	3.8	467
Gm-c1009	'Williams'	Root, 2-month-old	2 553	269	1 683	76	3.2	560
Gm-c1010	'Williams'	Cotyledon, young	1 115	91	553	58	6.2	664
Gm-c1011	'Williams'	Cotyledon, immature	322	40	166	64	3.9	646
Gm-c1012	'Williams'	Apical shoot, 9–10 days old	2 200	228	1 620	84	2.5	571
Gm-c1013	'Williams'	Seedling, 2–3 weeks old	2 923	309	1 658	67	4.1	618
Gm-c1014	'Williams'	Leaf, 2–3 weeks old	2 289	568	816	60	2.6	518
Gm-c1015	'Williams 82'	Flower, mature	5 690	685	3 241	69	3.6	493
Gm-c1016	'Williams 82'	Flower, immature	8 847	1 175	4 268	62	3.9	522
Gm-c1017	'Williams 82'	Vegetative bud	2 838	313	1 706	71	3.6	556
Gm-c1018	'Williams 82'	Leaf, 2–3 weeks old	1 223	125	829	78	3.2	506
Gm-c1019	'Williams'	Seed coat, immature	5 353	644	3 134	71	3.4	619
Gm-c1020	'Williams'	Nodule	1 139	85	829	80	3.6	498
Gm-c1023	T157	Seed coat	2 288	188	1 683	82	3.2	590
Gm-c1024	'Williams'	Hypocotyl, 9–10 days old	543	56	385	81	2.8	517
Gm-c1025	'Williams 82'	Hypocotyl, 3 days old	1 189	122	810	78	3.1	503
Gm-c1026	'Williams'	Leaf, senescing	600	49	427	79	3.5	474
Gm-c1027	'Williams'	Cotyledon, 3 and 7 days old	7 384	942	3 840	65	3.8	662
Gm-c1028	Supernod	Root, *B. japonicus*	8 950	1 339	4 511	65	3.3	702
Gm-c1029	'Williams'	Cotyledon, young	1 511	147	960	73	3.7	615
Gm-c1031	'Williams'	Seedling, no coty, 5 days old	1 603	193	914	69	3.6	556
Gm-c1032	'Williams'	Cotyledon, 8 days old	1 598	184	959	72	3.5	620
Gm-c1033	'Delsoy'	Root	1 345	144	881	76	3.2	547
Gm-c1034	'Williams'	Seed coat, young	773	41	553	77	5.4	584
Gm-c1035	'Williams'	Leaf, immature	1 782	262	969	69	3.1	531
Gm-c1036	'Jack'	Somatic embryo	1 434	116	1 167	89	2.3	563
Gm-c1037	'Williams'	Leaf, 2 weeks old	1 929	211	1 219	74	3.4	538
Gm-c1038	'Williams 82'	Leaf, senescing	823	115	491	74	2.9	610
Gm-c1039	'Ogden'	Seedling, no cotyledon	1 502	152	1 077	82	2.8	578
Gm-c1040	'Williams 82'	Germ seed, hyp and plu	1 698	170	1 244	83	2.7	646
Gm-c1041	'Williams 82'	Leaf, senescing	885	86	629	81	3.0	688
Gm-c1042	'Raiden'	Seedling, no cotyledon	1 573	159	1 048	77	3.3	634
Gm-c1043	'Williams'	Germ seed, hyp and plu	1 921	218	1 293	79	2.9	670
Gm-c1044	'Williams'	Hypocotyl, 9–10 days old	1 901	167	1 360	80	3.2	649
Gm-c1045	'Williams 82'	Hypocotyl, 9–10 days old	1 827	168	1 288	80	3.2	643
Gm-c1046	'Williams'	Germinating seed	1 147	103	827	81	3.1	657
Gm-c1047	'Williams'	Leaf, immature	829	83	512	72	3.8	482
Gm-c1048	'Clark'	Seedling, 1 week old	1 178	81	931	86	2.7	551
Gm-c1049	'Clark'	Seedling, 3 weeks old	1 904	162	1 415	83	3.0	602
Gm-c1050	'Clark'	Leaf, 3 weeks old	1 486	127	914	70	4.5	593
Gm-c1051	'Corolla'	Floral meristem	830	74	639	86	2.6	574
Gm-c1052	'Harosoy'	Seedling, 1 week old	1 082	96	636	68	4.6	602
Gm-c1061	'Raiden'	Flower, mature	1 550	104	1 153	81	3.8	612
Gm-c1062	'Raiden'	Stem, 1 month old	1 006	83	819	90	2.3	582
Average						**75**	**3.5**	**581**
Gm-r1021	'Williams'	Root, 8 days old	3 274	235	2 777	92	2.1	731
Gm-r1030	'Williams'	Cotyledon, mature	3 073	365	1 735	68	3.7	616
Gm-r1070		Mixed tissues	7 360	892	5 328	85	2.3	663
Average						**82**	**2.7**	**670**
Other			2 376					
Entire			121 068	16 928	17 336	28	6.1	788

Note: Size, contig size (mean no. of ESTs); Length, contig length (bp); hyp, hypocotyl; plu, plumule.

Fig. 1. Contig assembly information for individual libraries. Percent unigenes (A), contig size (B), and contig length (C) are indicated with respect to library size. ●, Gm-c libraries; ○, Gm-r libraries. Linear trendlines are indicated.

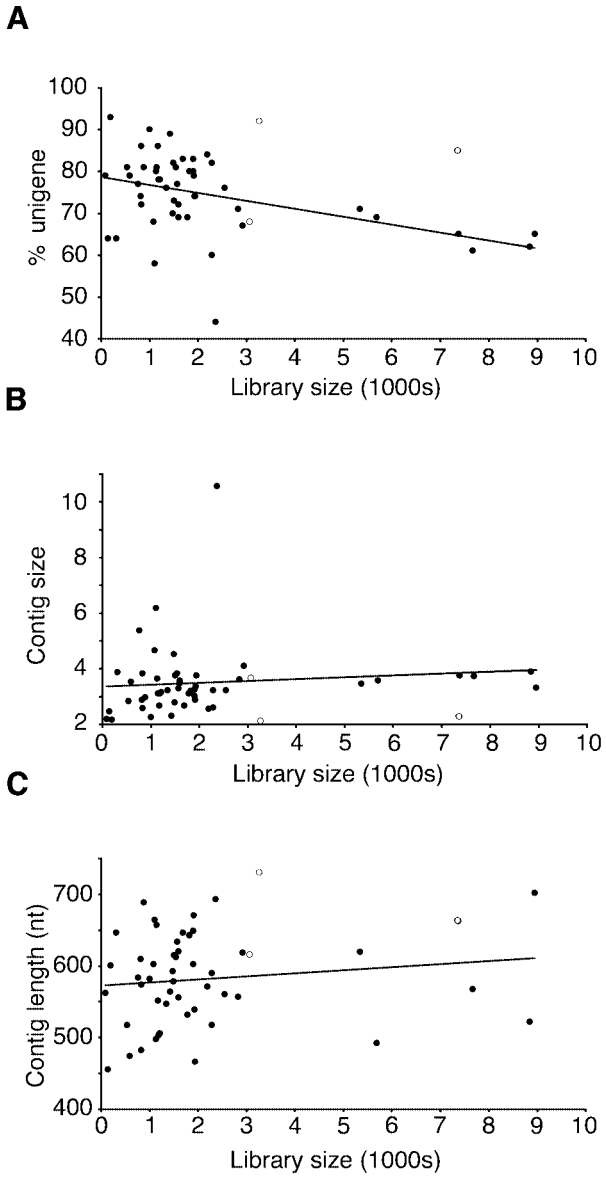
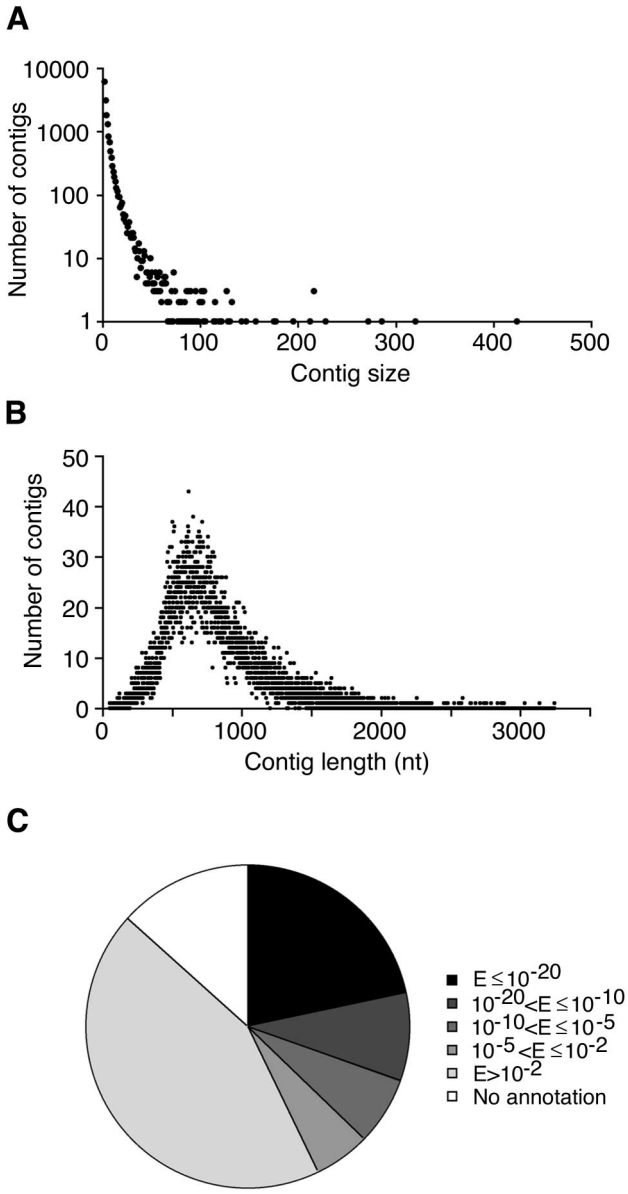


Fig. 2. Contig assembly information for the entire EST set. Distributions of contig size (A) and contig length (B) are indicated. Best matches for each unigene sequence were obtained from the SWISS PROT database. The percent distribution of *E* values of the best matches is indicated (C).



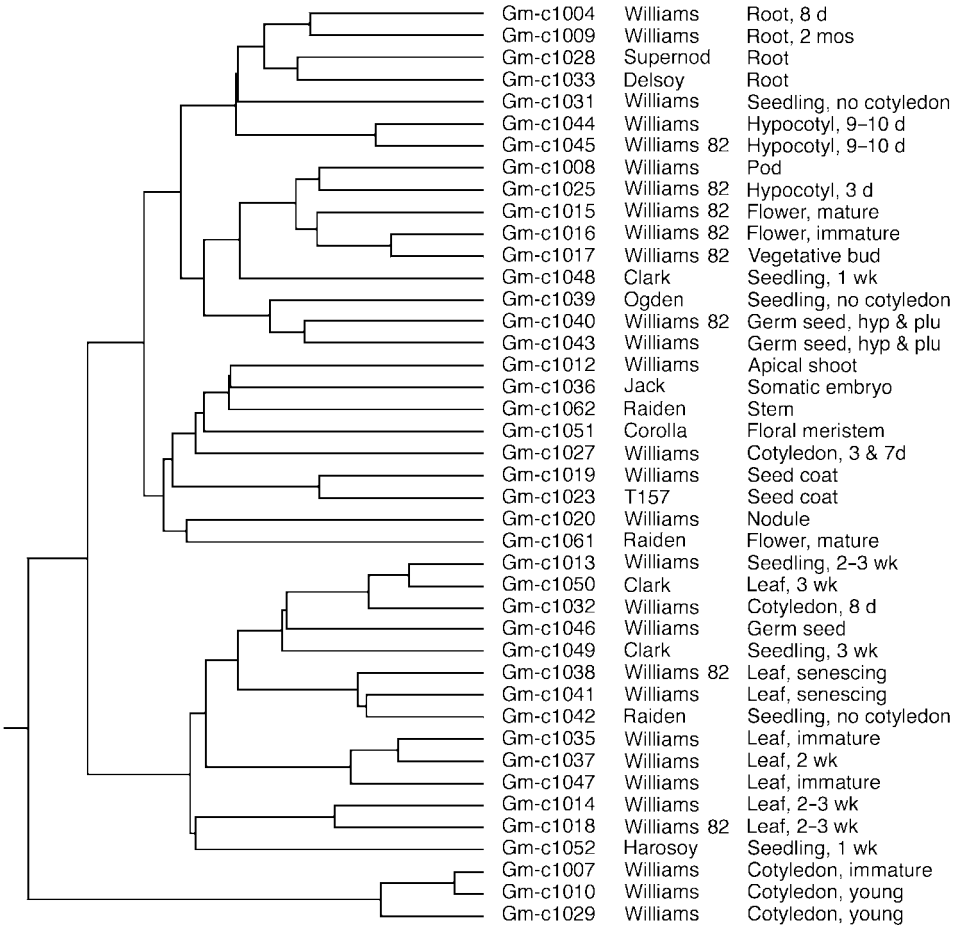
contigs consisted of more than 100 ESTs (Fig. 2A), whereas in some cases the contig length exceeded 3000 bp (Fig. 2B). Deduced amino acid sequences of the EST contigs were compared with sequences within SWISSPROT. Highly significant matches ($E \leq 10^{-10}$) occurred 30.7% of the time. A total of 13.4% of the sequences returned no matches (Fig. 2C). This fraction contains contigs of mostly untranslated mRNA as well as potentially novel genes.

Library clustering

After excluding libraries with less than 800 ESTs and all contigs with less than 10 ESTs, a total of 1763 contigs were used to cluster 42 libraries with respect to their expression profiles (Fig. 2). For each contig, the number of ESTs from each library was tabulated to create an expression profile. The numbers of ESTs comprising this contig–library matrix were then converted to percentages of each library. A Pearson correlation matrix of pairwise library–library comparisons was generated and the values from this matrix were used to generate a pairwise Euclidean distance matrix. Finally, UPGMA analysis was used to create a dendogram comparing the library expression patterns (Fig. 3).

In this analysis, libraries with similar EST composition and levels appear together on the dendogram. In most cases, libraries generated from similar organs tended to cluster together, i.e., root libraries and seed coat libraries, regardless of the genotype from which they were derived (Fig. 3). In other cases, such as with young cotyledon libraries, different genotypes seemed to present very different expression profiles. For example, in the case of flower-related libraries,

Fig. 3. Dendrogram showing relationships between libraries. Libraries were clustered with respect to similarity in contig content. Library names (first column), cultivars (second column), and tissue (third column) are indicated.



those from the cultivar Williams 82 clustered together within the same clade, whereas libraries from the cultivars Corolla and Raiden were distinctly separate. Leaf libraries also tended to cluster together with some exceptions. Seedling libraries from the cultivars Williams and Williams 82 had similar expression profiles, whereas seedling libraries Gm-c1048, -c1049, and -c1050, generated from a near-isogenic derivative of the cultivar Clark representing a unique compilation of gene combinations (Shoemaker and Specht 1995), presented library expression profiles distinct from other leaf and seedling libraries. These differences were also apparent among developmental stages. An immature flower library and a vegetative bud library from the cultivar Williams 82 were more closely related to each other than to a mature flower library from the same cultivar. In addition, a flower meristem library from the genotype 'Corolla' and a mature flower library from the cultivar Raiden presented very different expression profiles from each other and from the 'Williams 82' libraries (Fig. 3).

Two-dimensional representations of gene and library clustering

To combine clustering data for contigs and libraries, a pairwise contig–contig Pearson correlation matrix was generated as above. Euclidean distances were then generated for all contig comparisons and these distances were used to generate dendograms showing relatedness of expression profiles of contigs. The two dendograms, contig–contig and library–library, were then used to reorder the original percentage-based matrices so that libraries and contigs with similar expression profiles grouped together. Values were normalized with respect to contig size to facilitate visualization. This resulted in a two-dimensional qualitative representation of contig and library similarities based on expression profiles (Fig. 4).

Not surprisingly, seed composition-related genes, such as seed storage protein genes, clustered together. An examination of the two-dimensional representation shows that the highest level of expression of these genes is found in only a few libraries (Gm-c1007, -c1010, and -c1029) corresponding to cotyledon tissue. These libraries were also clustered, indicating that they had similar global patterns of gene expression. Other examples of correlated expression with specific libraries were seen. For example, photosynthesis-related genes were found mainly among libraries whose cDNAs were derived from green tissues, whereas nodulin and leghemoglobin genes were found only in tissue from soybean nodules (Fig. 4).

The predicted most highly expressed genes, based on their average percentage in all libraries, are shown in Fig. 5A. As expected, several photosynthetic genes as well as Rubisco small-subunit genes were found in high levels. Interestingly, various isoforms of the Rubisco genes were observed. Their expression profiles among different libraries suggest that dif-

Fig. 4. Grayscale diagram showing clustered expression patterns. Contigs were clustered according to similarity in expression profiles (vertical dendrogram). Selected contig annotations are indicated. Libraries were clustered according to similarity in contig content (horizontal dendrogram). White indicates 100% expression of a contig in a given library, black indicates 0% expression.

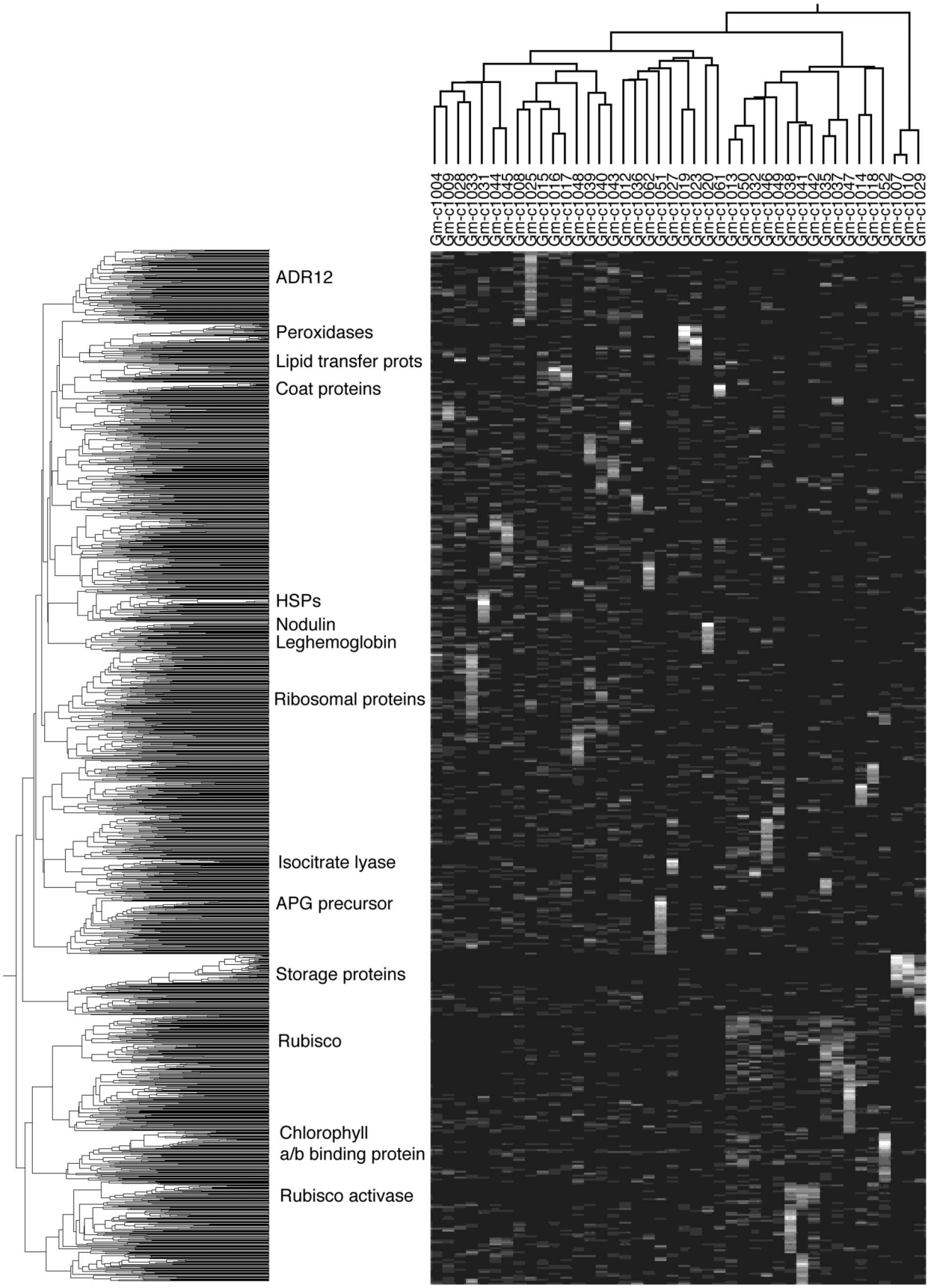
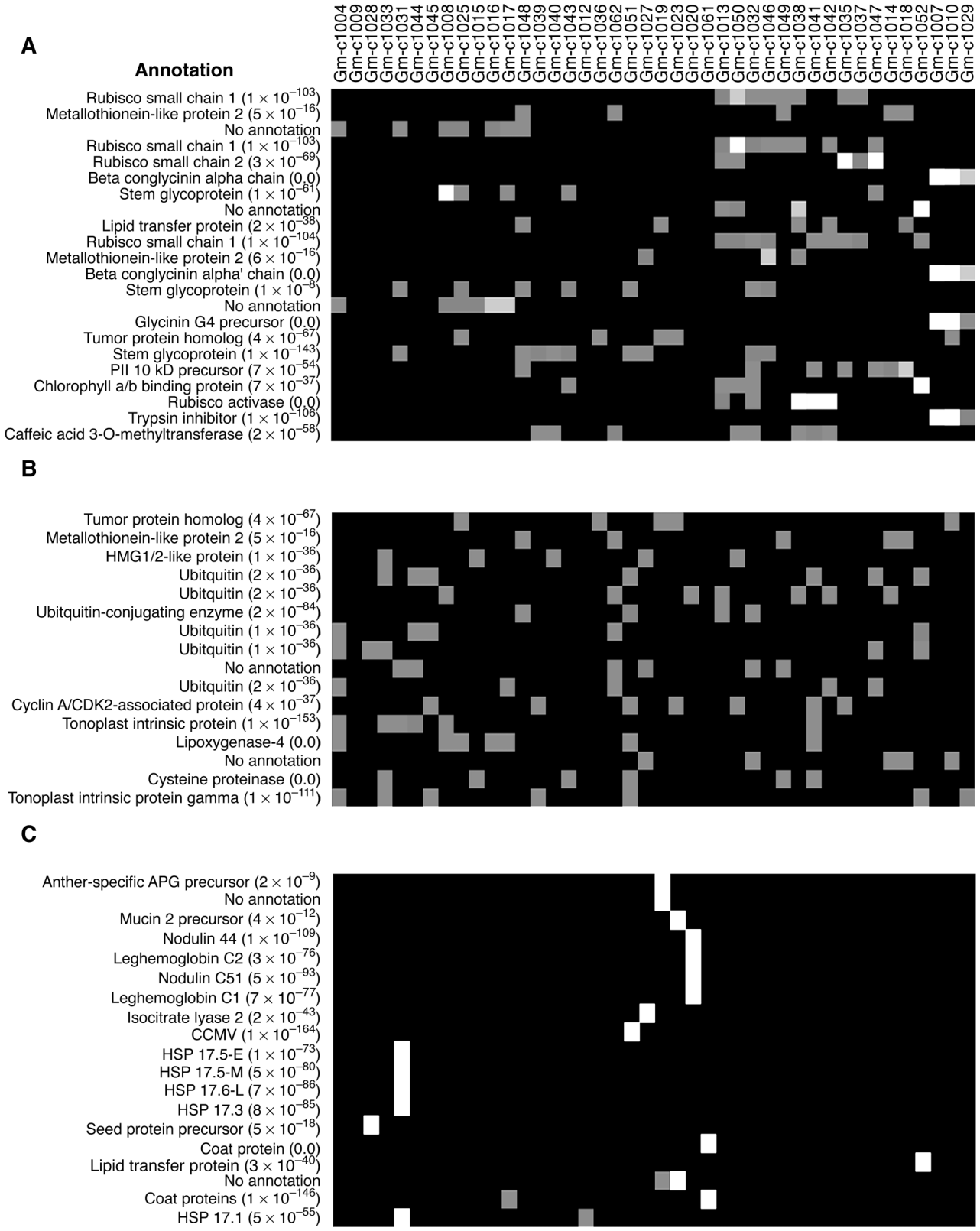


Fig. 5. Grayscale diagram showing the most highly expressed (A), most uniformly expressed (B), and most nonuniformly expressed (C) contigs. White indicates 100% expression of a contig in a given library, black indicates 0% expression. Annotations are based on the best matches for contig consensus sequences in the SWISS PROT database. *E* values are indicated in parentheses.



ferential expression may occur among members of this family. Metallothionein-like genes and stem glycoprotein genes also showed differential expression patterns among family members. Other highly expressed genes were observed only in a single isoform. For example, caffeic acid 3-*O*-methyltransferase was found to be highly expressed in a breadth of material ranging from seedling to senescing tissue. This gene has recently been implicated in the process of lignification in legumes (Guo et al. 2001).

These analyses also allowed for the identification of genes that are uniformly expressed across a wide range of libraries. Not surprisingly, ubiquitin isoforms are found in this category (Fig. 5B). Some ESTs were observed at relatively high levels in only one or a few specific libraries (Fig. 5C). For example, heat-shock gene transcripts were highly abundant in library Gm-c1031. This library was derived from whole seedlings of the cultivar Williams, minus cotyledons, that were heat shocked for 1 h at 40°C. Leghemoglobin and nodulin gene transcripts were observed at high levels in library Gm-c1020 (root nodule library). A putative anther-specific APG precursor gene product appeared at high levels in library Gm-c1019, a library made from seed coats of immature 200–300 mg seeds. However, the low *E* value associated with this annotation (2.0×10^{-9}) suggests that the gene product may actually represent a similar but unrelated proline-rich protein. Isocitrate lyase, a lipid mobilization enzyme normally detectable during late stages of embryo development, was identified in library Gm-c1027 (young cotyledons) with high *E* values (2.0×10^{-43}). Viral coat protein isoforms made up a significant percentage in libraries Gm-c1016 and -c1061, cDNA libraries made from mature flower of the cultivars Williams 82 and Raiden, respectively. Both samples were taken from field-grown plants and it is possible that these isoforms are present because of viral contamination or infection of the materials.

Examination of EST profiles across a wide range of organs, genotypes and environmental conditions may provide important information for analysis of gene function. Overlapping patterns of expression provide clues about the interrelationships of genes and gene groups. However, several conditions must be considered before relying upon EST expression profiles for this purpose. First, it must be assumed that each library is produced in a similar fashion and represents an unbiased and representative sampling of the transcripts present in the source tissue. Second, the ESTs must be assumed to represent a random sampling of the cDNAs present in the library. And third, enough redundancy must be present among the ESTs to provide meaningful estimates of levels of expression.

The identification of isoforms of many gene products probably reflects the ancient polyploidy of the soybean. It is thought that as many as 70% of all plant species are polyploid in origin, including the model plant species *Arabidopsis* (Grant et al. 2000; Vision et al. 2000). Analysis of RFLP patterns suggest that more than 90% of the non-repetitive sequences in soybean are present in two or more copies and the duplicated regions of soybean are readily identifiable by examination of hybridization-based genetic maps (Shoemaker et al. 1996).

The two-dimensional clustering of genes by expression profile presented here identifies genes that act in concert as well as the tissues in which they act. This method allows the identification of co-regulated genes that may not be intuitively recognized as such without this type of data analysis. With increasing numbers of ESTs representing a wide range of source tissue being deposited into public databases daily, it may become possible to conduct comparative analyses between ESTs of different species. These comparisons would provide insight into inter-specific gene evolution at the structural and functional levels.

Acknowledgements

Special acknowledgements are given to the Washington University EST Sequencing Group; Marilyn Gibbons, Erika Ritter, Catherine Franklin, Jennifer Bennett, Rusadan Tsagareishvili, Larisa Belaygorod, Irina Ronko, Lenon Maguire, Andrew Grow, Sara Kennedy, Lisa Marie Carr, and Mary Fedele. The authors are grateful for the funding and support of the North Central Soybean Research Program and the United Soybean Board. This research was funded in part by National Science Foundation Plant Genomics Grant No. DBI-9872565. Names are necessary to report factually on the available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable. Contribution of the Field Crops Research Unit, USDA–ARS, Midwest Area and Project No. 3236 of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA 50011. Journal paper No. 19463.

References

Adams, M., Kerlavage, A., Fleischmann, R., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature (London), **377**: 163–174.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polumeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science (Washington, D.C.), **252**: 1651–1656.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Arumuganathan, K., and Earle, E.D. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9**: 208–219.

Clemente, T., LaVallee, B., Howe, A., Ward, D-C., Rozman, R., Hunter, P., Broyles, D., Kasten, D., and Hinchee, M. 2000. Progeny analysis of glyphosate selected transgenic soybeans derived from Agrobacterium-mediated transformation. Crop Sci. **40**: 797–803.

Covitz, P.A., Smith, L.S., and Long, S.R. 1998. Expressed sequence tags from a root-hair-enriched *Medicago truncatula* cDNA library. Plant Physiol. **117**: 1325–1332.

Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker R.C., Lark, K.G., Kahler, A.L., Kaya, N., VanToai, T.T., Lohnes, D.G., Chung, J., and Specht, J.E. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. **39**: 1464–1490.

Delseny, M., Cooke, R., Raynal, M., and Grellet, F. 1997. The *Arabidopsis thaliana* cDNA sequencing projects. FEBS Lett. **405**: 129–132.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8**: 175–185.

Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S., and Claverie, J.-M. 1999. Large-scale statistical analysis of rice ESTs reveal correlated patterns of gene expression. Gen. Res. **9**: 950–959.

Felsenstein, J. 1989. PHYLIP – Phylogeny Inference Package version 3.2). Cladistics, **5**: 164–166.

Fernandes, J., Brendel, V., Gai, X., Lal, S., Chandler, V.L., Elumalai, R., Galbraith, D.W., Pierson, E., and Walbot, V. 2002. EST and microarray analysis of maize gene expression. Plant Physiol. In Press.

Goldberg, R.B. 1978. DNA sequence organization in the soybean plant. Biochem. Genet. **16**: 45–68.

Grant, D., Cregan, P., and Shoemaker, R. 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc. Natl. Acad. Sci. U.S.A. **97**: 4168–4173.

Guo, D., Chen, F., Inoue, K., Blount, J. and Dixon, R. 2001. Down regulation of caffeic acid 3-*O*-methyltransferase and caffeoyl coA 3-*O*-methyltransferase in transgenic alfalfa: impacts on lignin structure and implications for the biosynthesis of g and s lignin. Plant Cell, **13**: 73–78.

Gurley, W.B., Hepburn, A.G., and Key, J.L. 1979. Sequence organization of the soybean genome. Biochim. Biophys. Acta, **561**: 167–183.

Hatey, F., Tosser-Klopp, G., Clouscard-Martinato, C., Mulsant, P., and Gasser, F. 1998. Expressed sequence tags for genes: a review. Genet. Sel. Evol. **30**: 521–541.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B. et al. 1996. Generation and analysis of 280 000 human expressed sequence tags. Genome Res. **6**: 807–828.

Huang, X., and Madan, A. 1999. CAP3: A DNA sequence assembly program. Genome Res. **9**: 868–877.

Lim, C.O., Kim, H.Y., Kim, M.G., Lee, S.I., Chung, W.S., Park, S.H., Hwang, I., and Cho, M.J. 1996. Expressed sequence tags of chinese cabbage flower bud cDNA. Plant Physiol. **111**: 577–588.

Lin, X., Kaul, S., Rounsley, S., Shea, T., Benito, M.-I. et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature (London), **402**: 761–768.

Manger, I.D., Hehl, A., Parmley, S., Sibley, L.D., Marra, M., Hillier, L., Waterston, R., and Boothroyd, J.C. 1998. Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: identification of developmentally regulated genes. Infect Immunol. **66**: 1632–1637.

Marek, L.F., Mudge, J., Darnielle, L., Grant, D., Hanson, N., Paz, M., Huihuang, Y., Denny, R., Larson, K., Foster-Hartnett, D., Cooper, A., Danesh, D., Larsen, D., Schmidt, T., Staggs, R., Crow, J.A., Retzel, E., Young, N.D., and Shoemaker, R.C. 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome, **44**: 572–581.

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R. et al. 1999*a*. An encyclopedia of mouse genes. Nat. Genet. **21**: 191–194.

Marra, M.A., Kucaba, T.A., Hillier, L.W., and Waterston, R.H. 1999*b*. High-throughput plasmid DNA purification for 3 cents per sample. Nucleic Acids Res. **27**: 37e.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G. et al. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature (London), **402**: 769–777.

Putney, S.C., Herlihy, W.C., and Schimmel, P. 1983. A new tropin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature (London), **302**: 718–721.

Riley, P. 1999. USDA expects record soybean supply. Inform, **10**: 503–506.

Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M. et al. 2000. Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. Nature (London), **408**: 820–822.

Shoemaker, R.C. 1999. Soyabean genomics from 1985 to 2002. AgBiotechNet, **1**: 1–4.

Shoemaker, R.C., and Specht, J.E. 1995. Integration of the Soybean Molecular and Classical Genetic Linkage Groups. Crop Sci. **35**: 436–446.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., and Boerma, H.R. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics, **144**: 329–338.

SoyStats: A reference guide to important soybean facts and figures. 2000. American Soybean Association, St. Louis, Mo. Available from http://www.unitedsoybean.org/soystats2000/index.htm.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T.T., Boerjan, W., Gustafsson, P., Uhlen, M., Sundberg, B., and Lundeberg, J. 1998. Gene discovery in the wood-forming tissues of popular: analysis of 5,692 expressed sequence tags. Proc. Natl. Acad. Sci. U.S.A. **95**: 13 330 – 13 335.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T. et al. 2000. Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature (London), **408**: 823–826.

Tanabe, K., Nakagomi, S., Kiryu-Seo, S., Namikawa, K., Imai, Y., Ochi, T., Tohyama, M., and Kiyama, H. 1999. Expressed-sequence-tag approach to identify differentially expressed genes following peripheral nerve axotomy. Mol. Brain Res. **64**: 34–40.

Theologis, A., Ecker, J.R., Palm, C.J., Federspiel, N.A., Kaul, S. et al. 2000. Sequence and analysis of chomosome 1 of the plant *Arabidopsis thaliana*. Nature (London), **408**: 816–820.

Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis*. Science (Washington D.C.), **290**: 2114–2117.

Xing, A., Zhang, Z., Sato, S., Staswick, P., and Clemente, T. 2000. The use of the two T-DNA binary system to derive marker-free transgenic soybeans. In Vitro Cell. Dev. Biol.-P, **36**: 456–463.

Zhang, Z.A., Xing, P., Staswick, T., and Clemente, T. 1999. The use of glufosinate as a selective agent in Agrobacterium-mediated transformation of soybean. Plant Cell Tissue Organ Cult. **56**: 37–46.